

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-085615

(43)Date of publication of application : 30.03.1999

(51)Int.Cl.

G06F 12/08

G06F 15/163

G06F 15/173

(21)Application number : 09-243127

(71)Applicant : CANON INC

(22)Date of filing : 08.09.1997

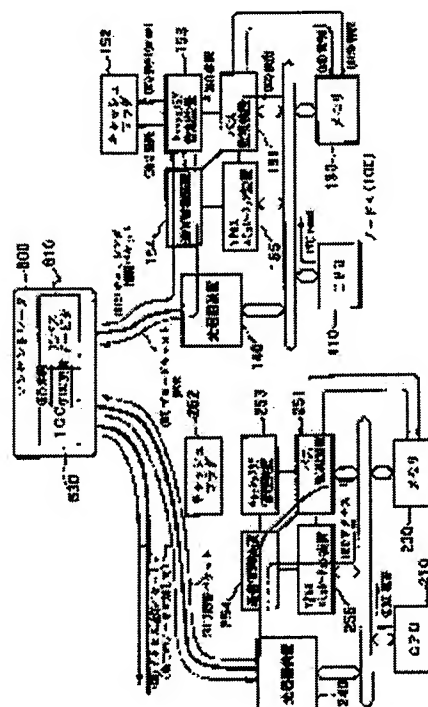
(72)Inventor : SHIMOYAMA TOMOHIKO

(54) SYSTEM AND DEVICE FOR PROCESSING INFORMATION AND CONTROL METHOD THEREFOR

(57)Abstract:

PROBLEM TO BE SOLVED: To reduce the capacity of a storage mechanism for recording the destination to multicast access information required for providing a distributed shared memory.

SOLUTION: Nodes A and B for caching data provided by access to a shared memory space hold cache flags 152 and 252 showing whether their own memory spaces are cached in the other nodes or not for the unit of a cache. When the access to a memory in the present node is detected, the cache flag is referred to and when the relevant access address is cached by the other node, the multicast of the relevant access is requested to a concentrator 600. An ICC 630 of the concentrator 600 holds the cache of more detailed directory information and in the case of cache hit, the relevant access is reported to the node shown by this directory information.



2 DI

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-85615

(43) 公開日 平成11年(1999) 3月30日

(51) Int.Cl.⁶

G 0 6 F 12/08
15/163
15/173

識別記号

3 1 0

F I

G 0 6 F 12/08 3 1 0 B
15/16 3 2 0 K
4 0 0 N

審査請求 未請求 請求項の数14 O.L (全 17 頁)

(21) 出願番号 特願平9-243127

(22) 出願日 平成9年(1997) 9月8日

(71) 出願人 000001007

キヤノン株式会社

東京都大田区下丸子3丁目30番2号

(72) 発明者 下山 朋彦

東京都大田区下丸子3丁目30番2号 キヤ
ノン株式会社内

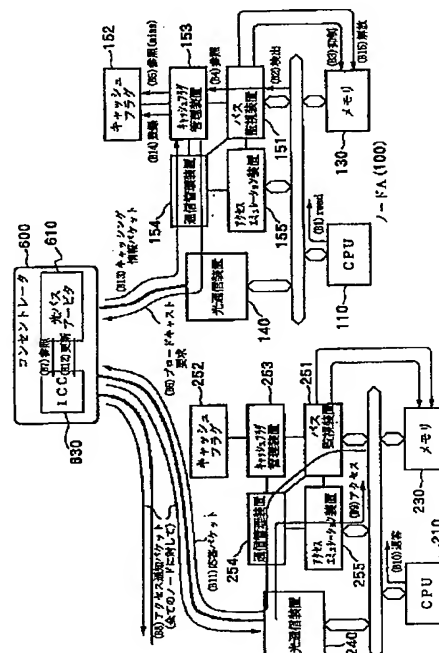
(74) 代理人 弁理士 大塚 康徳 (外2名)

(54) 【発明の名称】 情報処理システム及び情報処理装置及びその制御方法

(57) 【要約】

【課題】 分散共有メモリを実現する際に必要となるアクセス情報のマルチキャストの宛て先を記録するための記憶機構の容量を削減する。

【解決手段】 共有メモリ空間へのアクセスによって得られたデータをキャッシングするノードA、Bは、自身のメモリ空間に関して他のノードにキャッシングされているか否かをキャッシュ単位で示すキャッシュフラグ152、252が保持されている。自ノード内のメモリに対するアクセスが検出されるとキャッシュフラグを参照し、当該アクセスアドレスが他ノードにキャッシングされている場合は、コンセントレータ600に当該アクセスのマルチキャストを要求する。コンセントレータ600のICCB630にはより詳細なディレクトリ情報のキャッシュが保持されており、キャッシュヒットした場合は、このディレクトリ情報で示されるノードへ当該アクセスが通知される。



【特許請求の範囲】

【請求項1】 複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムであって、

前記共有メモリ空間へのアクセスによって得られたデータをアクセス元の情報処理装置にてキャッシングするキャッシュ手段と、

前記複数の情報処理装置の各々が、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持手段と、

前記保持手段によって保持されているキャッシュ情報に基づいて前記キャッシュ手段のキャッシュメンテナンスを行うメンテナンス手段とを備えることを特徴とする情報処理システム。

【請求項2】 前記メンテナンス手段は、前記複数の情報処理装置の各々において、自身のメモリ空間へのアクセスが発生した場合に、当該アクセス先アドレスが他の情報処理装置にキャッシングされているか否かを前記キャッシュ情報を参照して判定する判定手段と、

前記判定手段によって前記アクセス先アドレスが他の情報処理装置によってキャッシングされていると判定された場合、当該アクセスをマルチキャストするマルチキャスト手段とを備えることを特徴とする請求項1に記載の情報処理システム。

【請求項3】 前記複数の情報処理装置の一つが、キャッシングアドレスとキャッシング先の情報処理装置を示すマルチキャスト情報を保持する第2保持手段を備える通信管理装置であり、

前記マルチキャスト手段による前記アクセスのマルチキャストは、前記第2保持手段に保持されたマルチキャスト情報に基づいて選択された情報処理装置に対して行なわれることを特徴とする請求項2に記載の情報処理システム。

【請求項4】 前記マルチキャスト手段は、前記判定手段によって前記アクセス先アドレスが他の情報処理装置によってキャッシングされていると判定され、前記第2保持手段に対応する情報が保持されていない場合、当該アクセスを前記複数の情報処理装置の全てにマルチキャストすることを特徴とする請求項3に記載の情報処理システム。

【請求項5】 前記複数の情報処理装置を接続する通信網が、前記通信管理装置を中心としたスター結合であることを特徴とする請求項3に記載の情報処理システム。

【請求項6】 前記複数の情報処理装置を接続する通信網が、複数の波長の光を用いて接続する光波長多重化した経路により構成されることを特徴とする請求項1乃至5のいずれかに記載の情報処理システム。

【請求項7】 複数の情報処理装置が通信可能に接続さ

れ、相互に共有可能な共有メモリ空間を有する情報処理システムにおける情報処理装置であって、

前記共有メモリ空間へのアクセスによって得られたデータをキャッシングするキャッシュ手段と、

自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持手段と、

前記メモリ空間にアクセスが発生した場合、前記保持手段によって保持されているキャッシュ情報に基づいて前記キャッシュ手段のキャッシュメンテナンスを行うメンテナンス手段とを備えることを特徴とする情報処理装置。

【請求項8】 前記メンテナンス手段は、前記複数の情報処理装置の各々において、自身のメモリ空間へのアクセスが発生した場合に、当該アクセス先アドレスが他の情報処理装置にキャッシングされているか否かを前記キャッシュ情報を参照して判定する判定手段と、

前記判定手段によって前記アクセス先アドレスが他の情報処理装置によってキャッシングされていると判定された場合、当該アクセスのマルチキャストを要求する要求手段とを備えることを特徴とする請求項7に記載の情報処理装置。

【請求項9】 複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムの制御方法であって、

前記共有メモリ空間へのアクセスによって得られたデータをアクセス元の情報処理装置にてキャッシングするキャッシュ工程と、

前記複数の情報処理装置の各々が、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持工程と、

前記保持工程によって保持されたキャッシュ情報に基づいて前記キャッシュ工程のキャッシュメンテナンスを行うメンテナンス工程とを備えることを特徴とする制御方法。

【請求項10】 前記メンテナンス工程は、前記複数の情報処理装置の各々において、自身のメモリ空間へのアクセスが発生した場合に、当該アクセス先アドレスが他の情報処理装置にキャッシングされているか否かを前記キャッシュ情報を参照して判定する判定工程と、

前記判定工程によって前記アクセス先アドレスが他の情報処理装置によってキャッシングされていると判定された場合、当該アクセスをマルチキャストするマルチキャスト工程とを備えることを特徴とする請求項9に記載の制御方法。

【請求項11】 前記マルチキャスト工程は、前記判定工程によって前記アクセス先アドレスが他の情報処理装

10

20

30

40

50

置によってキャッシングされていると判定された場合、当該アクセスを前記複数の情報処理装置の全てにマルチキャストすることを特徴とする請求項9に記載の制御方法。

【請求項12】 前記複数の情報処理装置の一つが、キャッシングアドレスとキャッシング先の情報処理装置を示すマルチキャスト情報をメモリに保持する第2保持工程を備える通信管理装置であり、前記マルチキャスト工程による前記アクセスのマルチキャストは、前記第2保持工程によって前記メモリ保持されたマルチキャスト情報に基づいて選択された情報処理装置に対して行なわれることを特徴とする請求項11に記載の制御方法。

【請求項13】 複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムにおける情報処理装置の制御方法であって、前記共有メモリ空間へのアクセスによって得られたデータをキャッシングするキャッシュ工程と、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持工程と、前記メモリ空間にアクセスが発生した場合、前記保持工程によって保持されているキャッシュ情報に基づいて前記キャッシュ工程におけるキャッシュメンテナンスを行うメンテナンス工程とを備えることを特徴とする制御方法。

【請求項14】 複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムに適用可能な情報処理装置のための制御プログラムを格納する記憶媒体であって、該制御プログラムがコンピュータを、前記共有メモリ空間へのアクセスによって得られたデータをキャッシングするキャッシュ手段と、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持手段と、前記メモリ空間にアクセスが発生した場合、前記保持手段によって保持されているキャッシュ情報に基づいて前記キャッシュ手段のキャッシュメンテナンスを行うメンテナンス手段として機能させることを特徴とする記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は分散共有メモリとその上でのキャッシュ機構を有する情報処理システム及び情報処理装置及びその制御方法に関するものである。

【0002】

【従来の技術】分散共有メモリと其上でのキャッシュ機構を有する情報処理システムが提案されている。一般に、分散共有メモリを実現する場合、共有メモリの一貫

性保持のために必要な通信を必要な計算機にのみ配送するため、あるアクセスに関しての情報をどの計算機に送るかを記録する必要があった。例えばディレクトリ方式の分散共有メモリのキャッシュシステムにおいては、メモリのキャッシュライン毎にどの計算機にアクセス情報をマルチキャストしなければならないかを記録している。

【0003】だがどの計算機にマルチキャストするかを示す宛て先の保持に使用するメモリの容量は、通信機構上、計算機の数により増大してゆき、その容量を押さえることが求められている。一般に、容量を押さえるための手法として、宛て先の情報を欠落させ、不必要な計算機にまでマルチキャストを行うことが行われることが多かった。しかしながら、この手法では、記憶容量を節約できる反面、不必要な計算機にマルチキャストを行うことによる性能の低下が避けられなかった。

【0004】これを補う技術として、頻繁に使用されるメモリアドレスについては詳しく宛て先を記録し、他のメモリアドレスに関しては宛て先の情報を欠落させる（特願平8-119268）、頻繁に使用されるメモリアドレスについては宛て先を記録し他のメモリアドレスに関してはブロードキャストを行う（特願平8-31022）、といった手法が本出願人によって提案されている。

【0005】

【発明が解決しようとする課題】しかしながら、上記のように宛て先をグループ化するなどしても、まだ大容量の記憶機構が必要となる局面が生じたり、宛て先情報の欠落をブロードキャストで補うことによる性能上の問題が生じたりして、その実現が現実的でなくなる場合があった。

【0006】本発明は、分散共有メモリを実現する際に必要となるアクセス情報のマルチキャストの宛て先を記録するための記憶機構の容量の削減と不要なマルチキャストによる性能低下とのバランスを向上させ、低コストで高性能な情報処理システム及び情報処理装置及びその制御方法を提供することを目的とする。

【0007】

【課題を解決するための手段】上記の目的を達成するための本発明の情報処理システムは以下の構成を備える。すなわち、複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムであって、前記共有メモリ空間へのアクセスによって得られたデータをアクセス元の情報処理装置にてキャッシングするキャッシュ手段と、前記複数の情報処理装置の各々が、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持手段と、前記保持手段によって保持されているキャッシュ情報に基づいて前記キャッシュ手段のキャッシュメンテナンスを行うメ

10

20

30

40

50

メンテナンス手段とを備える。

【0008】また、上記の目的を達成するための本発明の情報処理装置は以下の構成を備える。すなわち、複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムにおける情報処理装置であって、前記共有メモリ空間へのアクセスによって得られたデータをキャッシングするキャッシュ手段と、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持手段と、前記メモリ空間にアクセスが発生した場合、前記保持手段によって保持されているキャッシュ情報に基づいて前記キャッシュ手段のキャッシュメンテナンスを行うメンテナンス手段とを備える。

【0009】また、上記の目的を達成するための本発明の情報処理システムの制御方法は以下の工程を備える。すなわち、複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムの制御方法であって、前記共有メモリ空間へのアクセスによって得られたデータをアクセス元の情報処理装置にてキャッシングするキャッシュ工程と、前記複数の情報処理装置の各々が、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持工程と、前記保持工程によって保持されたキャッシュ情報に基づいて前記キャッシュ工程のキャッシュメンテナンスを行うメンテナンス工程とを備える。

【0010】また、上記の目的を達成する本発明の情報処理装置の制御方法は、複数の情報処理装置が通信可能に接続され、相互に共有可能な共有メモリ空間を有する情報処理システムにおける情報処理装置の制御方法であって、前記共有メモリ空間へのアクセスによって得られたデータをキャッシングするキャッシュ工程と、自身のメモリ空間に関して、他の情報処理装置にキャッシングされているか否かをキャッシュ単位で示すキャッシュ情報を保持する保持工程と、前記メモリ空間にアクセスが発生した場合、前記保持工程によって保持されているキャッシュ情報に基づいて前記キャッシュ工程におけるキャッシュメンテナンスを行うメンテナンス工程とを備える。

【0011】

【発明の実施の形態】以下、図面を参照して本発明の好適な一実施形態を詳細に説明する。

【0012】＜第1の実施形態＞図1は、本実施形態の情報処理システムの採用する並列計算機システムの構成を示す図である。計算機は複数のノード（100, 200, 300, 400, 500）からなり、各々のノードはコンセントレータ600を通じて光ファイバによりネットワークを構成している。コンセントレータは、ノード間の通信を管理する。

【0013】各ノードは通常の計算機であり、各々1個または複数のCPUを持つ。各々のノードはアドレス空間を共有している。図2は本実施形態によるメモリアドレス空間を説明する図である。図2に示されるように、一つのCPUからみえるアドレス空間には他のノードのアドレス空間が含まれる。なお、ノード間通信は、自分のアドレス上に見えている相手のメモリに対し、直接にデータを書き込むことで行う。

【0014】ノード（100, 200, 300, 400, 500）は光通信装置（140, 240, 340, 440, 540）を通じ、アービトレーション回線（160, 260, 360, 460, 560）、データ回線（170, 270, 370, 470, 570）の2本の光ファイバを介してコンセントレータ600に接続することによりネットワークを構成している。アービトレーション回線（160, 260, 360, 460, 560）は、ノード（100, 200, 300, 400, 500）とコンセントレータ600内の光バスアービタ610を接続する。アービトレーション回線により、各ノードは光バスアービタ610と相互に通信することができる。データ回線（170, 270, 370, 470, 570）は、ノードとコンセントレータ内のスターカブラ620を接続する。スターカブラ620は、一端に光を入力すると他端からその光が均等に出力される。従って、一つのノードがデータ光線に光を発すると、その光を他の全てのノードで受け取ることができる。この結果、一つのノードがデータ回線に発した光通信を、全てのノードで受け取ることが可能となる。

【0015】ノードは、通信する波長を選択することにより、任意のノードと通信を行う。例えば、ノードA100がノードB200と通信を行う場合は以下の手順が実行される。まず、ノードA100がアービトレーション回線160を通じて光バスアービタ610にノードB200との通信要求を送る。そして、光バスアービタ610はノードB200が他のノードと通信中でないことを確認すると、どのノードも使用していない波長 α をノードA, B間（100, 200間）の通信に割当てる。アービトレーション回線（160, 260）を通じてノードA, B（100, 200）に波長 α を使用して通信を行うように指示が出されると、ノードA, B（100, 200）が波長 α を使用してデータ回線（170, 270）を通じて通信を行う。

【0016】分散共有メモリは、このようなネットワークの上で光通信により実現される。例としてノードA100のプロセッサ110（以下、CPU110）が、ノードB200のメモリ230をリードする様子を図3に示す。なお、（A1）～（A7）に示す動作は一連のものとして実施される。

【0017】図3において、（A1）ノードA100のCPU110が、ノードBの

メモリ230に対してリードアクセスを発行する。

(A2) ノードA100の光通信装置140は、メインバス上のバスアクセスを監視している。メインバス上に、そのノードAのメモリ130以外のアドレスに対してアクセスが発行されると、光通信装置140はそのバスアクセスを検出する。

(A3) 光通信装置140は、アクセス要求バケットをコンセントレータ600を通じて、ノードBの光通信装置240に対して送る。

(A4) ノードBの光通信装置240はアクセス要求バケットを受け取ると、その依頼にしたがってノードBのメモリ230に対してのアクセスを代行する。

(A5) ノードBのメモリ230がアクセスに応答する。

(A6) ノードBの光通信装置240は、アクセスが終わると、ノードaの光通信装置140に対してアクノリッジを返す。

(A7) アクノリッジを受け取ったノードAの光通信装置140は、ノードBのメモリ(230)の代わりにノードAのCPU110のアクセスに応答する。

【0018】また本実施形態では、分散共有メモリ上で、ディレクトリ方式の一貫性保持方式を採用したキャッシュに類似したシステムを搭載している。ディレクトリ方式のキャッシュの詳細については、「共有記憶型並列システムの実例」(コロナ社刊)に示されている。

【0019】通常のディレクトリ方式のキャッシュ一貫性保持機構では、メモリの各キャッシュアドレス毎に(正確にはキャッシュ単位毎であるが、以下、これをキャッシュアドレス毎ということにする)そのアドレスをキャッシングしているノードを記録する。例えばノードがA~Hまでであるとすれば、各キャッシュアドレス毎に8bitのディレクトリを待つ。ディレクトリの各ビットは、ノードA~Hに対応する。もしノードA、B、F、Hにそのアドレスがキャッシングされていれば、そのキャッシュアドレスに対応したディレクトリの値は2進数で11000101となる。

【0020】本実施形態では上述の様なディレクトリ方式の代わりに、各キャッシュアドレス毎に1bitのフラグを持たせる。以降このフラグをキャッシングフラグと呼ぶ。キャッシングフラグは、そのキャッシュアドレスが他のノードにキャッシングされているかどうかを示す。例えばノードAのメモリ130内のアドレス18000000がノードA、B、Hにキャッシングされていれば、アドレス18000000に対応するキャッシングフラグはONとなる。キャッシュアドレス18000000がどのノードにもキャッシングされていないか、もしくはノードA自身のみキャッシングされている場合には、キャッシュアドレス18000000に対応するキャッシングフラグはOFFになる。

【0021】ノード内のCPU間でのキャッシュの一貫

性保持は、CPUに内蔵されたMESIプロトコル(Modified, Exclusive, Shared, Invalid)の4状態によりキャッシュを管理するプロトコル)によるスヌープキャッシュで管理されている(この実施形態ではMESIプロトコルを使用するが、これは本発明を限定するものではない)。

【0022】さて、各ノード間のキャッシュの一貫性保持は、外付けされたキャッシュ管理装置(150、250、350、450、550)により管理される。キャッシュ管理装置は内部バス上でアクセスを検出すると、そのアクセスアドレスに対応したキャッシングフラグを参照し、もしキャッシングフラグがONになっていたらバス上に発行されたアクセス(キャッシュメンテナンス情報)を他のノードに伝達する。

【0023】実際のキャッシュメンテナンス情報のマルチキャストは、コンセントレータ600内の光バスアービタ610により行われる。光バスアービタ610は、各ノードのキャッシュ管理装置からの要求によりアクセスを各ノードに伝達する。光バスアービタ610にはICC630と呼ばれるアクセスマルチキャスト情報のキャッシュがある。ICC630は最近使用されたアドレスについて、そのアドレスへのアクセスをどのノードに伝えるべきかを記録するキャッシュである。光バスアービタ610はアクセス要求を受け取ると、ICC630を参照してそのアクセスを伝達すべきノードが記録されていないかを調べる。もし記録されていたら(以下、ICCヒットという)、光バスアービタ610はそれらのノードにアクセスをマルチキャストする。もし記録されていなければ(以下、ICCミスという)、全てのノードにアクセスをブロードキャストする。

【0024】以下、キャッシュ管理装置及び光バスアービタの構成について説明する。

【0025】図4はノードAにおけるキャッシュ管理装置150の構成を示すブロック図である。ノード間のキャッシュの一貫性保持をするキャッシュ管理装置150は、バス監視装置151、キャッシングフラグ152、キャッシュフラグ管理装置153、通信管理装置154、アクセスエミュレーション装置155を備える。以下にキャッシュ管理装置150の備える各装置について説明する。

【0026】図5は、バス監視装置151の構成を示すブロック図である。バス監視装置151の内部は、シーケンサ151aとアドレスラッチ151bからなっている。シーケンサ151aはノード内のバスを監視し、バスマスタ(例えばCPU110)によるメモリ130のリードアクセス、invalidate、Read-with-Intent-to-Modify、メモリ書込みの実行を検出する。シーケンサ151aは、バス上のアクセス要求信号、バスアクセスに伴ってバスマスタから出力されるアクセス修飾信号などを監視することで、アクセスを検出する。

【0027】バス監視装置151はアクセス要求信号が有効であり、アクセスアドレスが自分のノードのメモリに対するものであったとき、もしくは他ノードのメモリへのキャッシュメンテナンスアクセスであったとき、アクセスを検出したと判断する。アクセスが自ノードのメモリに対するものであったとき、バス監視装置151はアクセスを検出するとメモリ130の応答を押さえ、その間にキャッシュフラグ管理装置153を通じてキャッシングフラグ152を調べる。もし対応するキャッシングフラグがONなら、バス監視装置151は光バスアービタ610にアクセスの伝達要求を出す。そしてそれらに対する応答により対象アドレスが他のノードにキャッシングされているかどうかを知り、キャッシュフラグ管理装置153を通じてキャッシュフラグ152を更新する。その後バス監視装置151がメモリ130の応答を許可することにより、CPU110はそのアクセスを完了する。一方、他ノードのメモリへのキャッシュメンテナンスアクセスであったときは、CPU110のアクセスをリトライさせておき、その間に光バスアービタ610にアクセスの伝達要求を出す。そして、その完了パケットが届いたらCPU110のアクセスリトライを解除し、処理を続行させる。

【0028】なお、上記メモリアクセスの中断は、バス上にアクセス再実行信号を出力し、そのメモリアクセスをプロセッサにリトライさせることで実現したが、他の方法によりアクセスを中断してもよい。

【0029】図6はキャッシュフラグ152及びキャッシュフラグ管理装置の構成を説明するブロック図である。先に述べたように本実施形態は、メモリの各キャッシュブロックに対して1bitのキャッシュフラグを持つ。キャッシングフラグは、そのアドレスが他のノードにキャッシングされているかどうかを示す。キャッシュフラグ152への操作は、キャッシュフラグ管理装置153により行われる。キャッシュフラグ管理装置153はシーケンサ153aにより管理されている。キャッシュフラグ152は、バス監視装置151が他ノードからのリード/ライトアクセスを検出した際に、対象メモリブロックが外部ノードにキャッシングされていることを記録する場合にONにされ、バス管理装置151がメモリ書込み/invalidata/Read-with-Intent-to-Modifyを

検出した場合、或いは対象メモリブロックが自ノードにのみキャッシングされている場合にはOFFにされる。

【0030】通信管理装置154はバス監視装置151、キャッシュフラグ管理装置153からの要求により、光通信装置140とのコミュニケーションを行う。通信管理装置154は、バス監視装置151、キャッシュフラグ管理装置153からのリード/invalidata/Read-with-Intent-to-Modify/メモリ書込みのマルチキャスト要求を受け、光通信装置140を介して光バスアービタ610に、アクセスのマルチキャスト/ブロードキ

ャストパケットを送出する。また光通信装置140を通じて受け取ったキャッシュの一貫性保持動作要求パケットにより、メモリアクセスエミュレート装置155に対してinvalidata/Read-with-Intent-to-Modify/ライト/リード要求を出力する。

【0031】図7はメモリアクセスエミュレート装置の構成を示すブロック図である。メモリアクセスエミュレート装置155は、通信管理装置154からの要求により、ダミーのメモリライトアクセス/ダミーのメモリリードアクセス/invalidata/Read-with-Intent-to-Modifyアクセスを自ノードのバス上に発行する。これらのアクセスにより、ノード間のキャッシュの一貫性を保持する。エミュレート装置155の発行するアクセスは、自分のノードに割り当てられたアドレスではありえないため、ノード内のバススレーブ（例えばメモリ130）より応答はありえず、ダミーのアクセスとなる。

【0032】次に、コンセントレータ600について説明する。図8はコンセントレータ600の構成を示すブロック図である。コンセントレータ600は、光バスアービタ610、スターカブラ620、ICC630を備える。光バスアービタ610は先に述べたように光回線の通信制御を行う。またノードから送られてきたマルチキャスト要求により、各ノードへのパケットの転送を行う。また、ICC630はアクセスのマルチキャスト先を保持するキャッシュメモリである。光バスアービタ610は、マルチキャスト対象アドレスがICC630にヒット（ICCヒット）した場合にはICC630に記録されたノードに対してマルチキャストを行う。ICC630にミス（ICCミス）した場合は全てのノードに対してブロードキャストを行う。

【0033】ICC630を用いることにより、マルチキャストの必要なノードを正確に知ることができるようになり、不必要なブロードキャストがなくなる。このため、当該情報処理システムの性能を向上することができる。図9はICC630のデータ構成を示す図である。本実施形態ではICC630は図9のようなフルマップ（各ノードにつき1ビットを割り当てる）方式をとっている。

【0034】なお、本実施形態では、光バスアービタ610は一つの計算機として実現するので、これらの機構はソフトウェアで実現されるものとする。しかしながら、この構成例は本発明を制限するものではなく、ハードウェアで実現することも可能であることは明らかである。

【0035】以上の様な構成を備える本実施形態の動作について以下に詳細に説明する。

【0036】図10はキャッシュ管理装置の動作を説明するフローチャートである。まず、ステップS101において、バス監視装置151がバスアクセスを検出するとステップS102へ進む。ステップS102では、当

該アクセスが自ノード内のメモリに対するアクセスか否かを判定する。そして、自ノード内のメモリへのアクセスであった場合はステップS103へ、そうでない場合はステップS111へそれぞれ進む。

【0037】ステップS103では、バス監視装置151がメモリ130の応答を抑制する。そして、キャッシュフラグ管理装置153を通してキャッシュフラグ152を参照し、当該キャッシュアドレスが他ノードにキャッシングされているか否かを判定する。他ノードへにキャッシングされていない場合は、そのままステップS108へ進み、メモリ130に対する応答抑制を解除し、当該アクセスを完了させる。

【0038】一方、ステップS104で当該キャッシュアドレスが他ノードに対してキャッシングされていると判定された場合は、他のノードに当該アクセスを通知するためにステップS105へ進む。ステップS105では、通信管理装置154に対して当該アクセスを通知するべく指示を出す。通信管理装置154は光通信装置140を通じて、コンセントレータ600にマルチキャスト要求を通知する。そして、コンセントレータ600よりマルチキャスト完了パケット（マルチキャスト完了パケットには、当該キャッシュアドレスが他ノードへキャッシングされている否かを示す情報が含まれる）を受信すると、このパケットに従ってキャッシュフラグ152を更新する。

【0039】その後、ステップS108にて、メモリ130に対する応答抑制を解除し、当該アクセスを完了させる。また、ステップS102で自ノードのメモリではないと判定された場合は、ステップS111へ進み、当該アクセスがキャッシュメンテナンスを含むか否かを判定する。キャッシュメンテナンス情報を含まない場合は、図3で説明した様な手順によって、光通信装置からアクセス要求パケットが発行されることになる。また、キャッシュメンテナンスを含むアクセスであれば、ステップS112へ進み、メモリをリトライさせる。そして、ステップS113において、対象メモリを有するノードに対してアクセス要求パケットを送出し、その応答を待つ。そして、応答を得たならば、ステップS114にてメモリのリトライを解除し、本アクセスを完了する。なお、ステップS111～S114に関しても、キャッシュ管理装置150内のバス管理装置151、通信管理装置154が処理している。

【0040】図11は、光バスアービタよりパケットを受信した場合の動作を説明するフローチャートである。光バスアービタ610よりパケットを受信すると、アクセスエミュレーション装置155は当該パケットに従ってバス上にアクセス情報を出力する（ステップS151、S152）。この結果、バスをスヌープしているCPU110、120やバス監視装置151が応答情報を生成し、通信管理装置154がこの応答情報をアクセス

結果として光バスアービタ610に通知する（ステップS153）。

【0041】図12は、光バスアービタの動作手順を説明するフローチャートである。図12では、キャッシュメンテナンスに係るマルチキャスト（ブロードキャストを含む）に対する応答動作が示されている。

【0042】ステップS201において、キャッシュアドレスを含むマルチキャスト要求パケットを受信すると、ステップS202へ進み、当該キャッシュアドレスでICC630を検索する。ICCミスした場合は、ステップS203へ進み、全ノードに対してマルチキャスト（ブロードキャスト）を行い、ステップS205へ進む。一方、ステップS202においてICCヒットした場合は、ステップS204へ進み、当該キャッシュアドレスに対応して記録されているノードに対してマルチキャストを行い、ステップS205へ進む。

【0043】ステップS205では、マルチキャスト先の各ノードからの完了パケットを待つ。各ノードからの完了パケットを受信したら、ステップS206へ進み、受信した完了パケットに従ってICC630を更新する。そして、ステップS207において、当該マルチキャスト要求もとのノードに対してマルチキャスト完了パケットを送出する。

【0044】次にキャッシュ管理装置150の動作を更に具体的に説明する。すなわち、

- ・ノード内のアドレスにリード、invalidate、Read-with-Intent-to-Modify（キャッシュのライトミス時に出力されるライトを前提としたリードサイクルであり、以下RWITMとする）などのキャッシュメンテナンス情報が出力されたとき、及び

- ・ノード外のアドレスに対してinvalidate、RWITMなどのキャッシュメンテナンス情報が出力されたときについて具体的な動作を説明する。

【0045】図13は、ノード内のメモリに対してリード/ライト/invalidate/RWITM等のバスアクセスが発行された場合の動作を説明する図である。

（B1）ノード内のメモリ130に対してリードアクセスが行われる。

（B2）バス監視装置151がそのアクセスを検知する。

（B3）メモリ130へのアクセスを検知したバス監視装置151は、メモリ130の応答を抑制する。

（B4）その間にバス監視装置151はキャッシュフラグ管理装置153に、当該キャッシュアドレスに関する参照要求を出す。

（B5）キャッシュフラグ管理装置153はキャッシュフラグ152に指定されたキャッシュアドレスが他ノードにキャッシングされているか（対象アドレスに対応したビットがONかOFFか）調べる。

（B6）もし対応するキャッシュフラグがONならば、

キャッシュフラグ管理装置153は通信管理装置154を通じて発行されたリードアクセスを他のノードにマルチキャストするよう光バスアービタ610に要求を出す。

(B7) 光バスアービタ610はICC630を参照する。

(B8) ICCミスならば光バスアービタ610は全てのノードに当該リードアクセスをブロードキャストする。

(B9) リードアクセスをブロードキャストされた各ノード(200, 300, 400, 500)は、バス上にリードアクセスを発行する。

(B10) 各ノードのCPU(210, 310, 410, 510)はこれをスヌープし、そのリードアクセスされたメモリ番地をキャッシングしているかどうかを返答する。

(B11) 各ノードのバス監視装置(251, 351, 451, 551)はこれを検出し、通信管理装置(254, 354, 454, 554)を通じて、アクセスが完了したこととそのノードで該当アドレスがキャッシングしているかどうかを光バスアービタ610に伝える。

(B12) 各ノードからリードアクセスが完了したことを通知された光バスアービタ610は、各ノードがキャッシングしている／していないという情報をICC630にキャッシングし、ICC630を更新する。

(B13) 他ノードに対象番地がキャッシングされているかどうかの情報を含んだアクセス完了バケットをノードA100に伝える。

(B14) 通信管理装置154を通じて各ノードのキャッシング状況を受け取ったキャッシュフラグ管理装置153は、その内容をキャッシュフラグ152に記録する。

(B15) 記録が終了すると、バス監視装置151はメモリ130の応答抑制を解除し、当該リードアクセスが完了する。

【0046】もし(B8)においてICCヒットならば、光バスアービタ610は記録されたノードにアクセスをマルチキャストする。その後の動作は上記の物と同様である。

【0047】上記手順によれば、最初の1回目のアクセスは、アクセス情報を必要のないノードに対してもブロードキャストを生じさせる。しかし、必要のないノードに対してのブロードキャストは、システムの性能を落とすことにはなるが論理的な矛盾を生じさせるものではない。各ノードがブロードキャストに応答して返ってくるバケットの中に、そのノードが該当アドレスをキャッシングしているかの情報が含まれているので、それに基づきキャッシュフラグをアップデートすることで、2回目以降のアクセスは必要なノードにのみマルチキャストが行われることになる。

【0048】また、ノード外のアドレスに対してリード／ライト／invalidata等のバスアクセスが出力された時は、光通信装置を通じて他ノードにアクセスを依頼する。アクセス先のノードのキャッシュ管理装置は、先に述べたように光通信装置を通じて行われたアクセスに対して応答する。

【0049】以上の様に、上記実施形態によれば、すべてのキャッシュアドレスについてそのアドレスが他の計算機にキャッシングされているかどうかを示すキャッシングフラグが記録され、他のノードにキャッシングされているアドレスに対してのメモリアccessのみが他の計算機に伝達されることになる。このため、効率のよいマルチキャスト要求が実現されることになる。

【0050】なお、以上のような構成をとるのは、計算機外への伝達コストが比較的高いためである。計算機外に伝達が必要なアクセスが多発した場合には、システム全体としての性能を発揮することができない。そこで外部へのアクセスであるかどうかを的確に判定し、計算機外への不要なアクセスを無くす或いは低減することでシステムの性能の向上が図られる。

【0051】<第2の実施形態>次に本発明の第2の実施形態を、図面を参照して説明する。

【0052】第1の実施形態ではICC630の構造を、フルマップ構造(ひとつのノードを1ビットで示す方法)で記録するようになっていた。このような場合、システムのノード数が多くなればICC630が必要とするメモリ容量も増加する。また、メモリ容量の増大化を防止すれば、その分登録可能なキャッシングアドレスが減少してしまう。そこで第2の実施形態はICC630の構成を変更し、1つのキャッシングアドレスに関する記憶容量の削減を図る。

【0053】図14は、第2の実施形態におけるICC630の構造を示す図である。ICC630の各データレコードは、キャッシュタグ630a、ノードグループ630b、マップ情報630cで構成される。キャッシュタグ630aはキャッシングアドレスを示す。ノードグループ630bには、ノードグループ内の先頭のノードが格納される。本例では、ノードはAからZまでの順に順序付けがなされており、ノードグループ630bによって先頭のノードとして指定されたノードから順に6つのノードがノードグループとなる。例えば、ノードグループ630bにノードAが指定されていれば、ノードAからノードFがノードグループとして指定されたことになる。

【0054】ICC630の下位はマップ情報630cであり、これは、当該グループ内のノードについてのフルマップの情報である。例えば上位でノードSが示されていた場合、下位ではノードS～Xについてのフルマップの情報が示される。同様に上位でノードCが示されていた場合、下位ではノードC～Hがフルマップで示され

ている。ノードAとJが同時に同じ番地をキャッシングした場合はこのような形式では対応できないので、IC630の上位を特別のパターン（例えば全て1）にして、その番地に対するアクセスはブロードキャスト（全てのノードに対して放送する）する。

【0055】以上の様に第2の実施形態によれば、グループ内で局所的なメモリ共有が行われている場合に、このような構成にすることでIC630の容量の削減を図ることができる。

【0056】以上説明したように、上記各実施形態によれば、各ノードに大容量のディレクトリ（メモリ）を備えることなく、フルマップ法式を採用したディレクトリ方式に匹敵する高性能のキャッシュシステムを構築することができる。

【0057】なお、本発明は、複数の機器（例えばホストコンピュータ、インタフェイス機器、リーダ、プリンタなど）から構成されるシステムに適用しても、一つの機器からなる装置（例えば、複写機、ファクシミリ装置など）に適用してもよい。

【0058】また、本発明の目的は、前述した実施形態の機能を実現するソフトウェアのプログラムコードを記録した記憶媒体を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはCPUやMPU）が記憶媒体に格納されたプログラムコードを読み出し実行することによっても、達成されることは言うまでもない。

【0059】この場合、記憶媒体から読出されたプログラムコード自体が前述した実施形態の機能を実現することになり、そのプログラムコードを記憶した記憶媒体は本発明を構成することになる。

【0060】プログラムコードを供給するための記憶媒体としては、例えば、フロッピーディスク、ハードディスク、光ディスク、光磁気ディスク、CD-ROM、CD-R、磁気テープ、不揮発性のメモリカード、ROMなどを用いることができる。

【0061】また、コンピュータが読出したプログラムコードを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているOS（オペレーティングシステム）などが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【0062】さらに、記憶媒体から読出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるCPUなどが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【0063】

【発明の効果】以上説明したように、本発明によれば、分散共有メモリを実現する際に必要となるアクセス情報のマルチキャストの宛て先を記録するための記憶機構の容量の削減と不要なマルチキャストによる性能低下とのバランスを向上させ、分散共有メモリシステムにおいて低コストで高性能なキャッシュシステムを提供できる。

【0064】

【図面の簡単な説明】

【図1】本実施形態の情報処理システムの採用する並列計算機システムの構成を示す図である。

【図2】本実施形態によるメモリアドレス空間を説明する図である。

【図3】ノードA100のCPU110が、ノードB200のメモリ230をリードする様子を説明する図である。

【図4】ノードAにおけるキャッシュ管理装置150の構成を示すブロック図である。

【図5】バス監視装置151の構成を示すブロック図である。

【図6】キャッシュフラグ152及びキャッシュフラグ管理装置の構成を説明するブロック図である。

【図7】メモリアクセスエミュレート装置の構成を示すブロック図である。

【図8】コンセントレータ600の構成を示すブロック図である。

【図9】IC630のデータ構成を示す図である。

【図10】キャッシュ管理装置の動作を説明するフローチャートである。

【図11】光バスアービタよりバケットを受信した場合の動作を説明するフローチャートである。

【図12】光バスアービタの動作手順を説明するフローチャートである。

【図13】ノード内のメモリに対してリード/ライト/invalidate/RWITM等のバスアクセスが発行された場合の動作を説明する図である。

【図14】第2の実施形態におけるIC630の構造を示す図である。

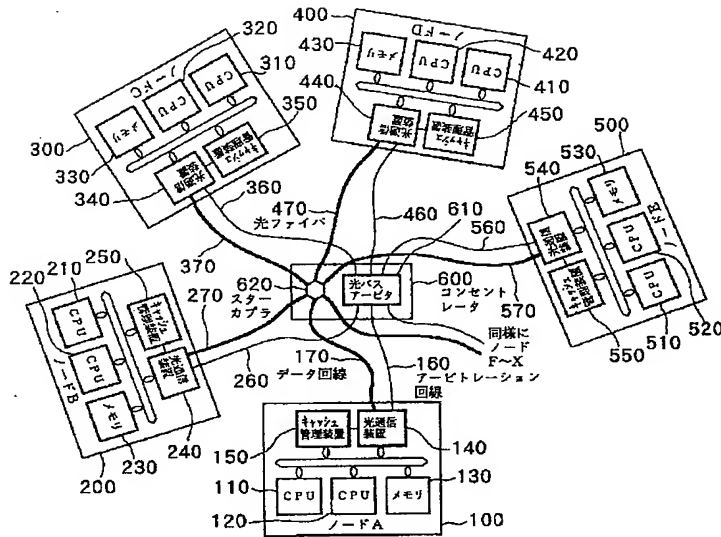
【符号の説明】

100, 200, 300, 400, 500	ノード
110, 120, 210, 220, 310, 320, 410, 420, 510, 520	CPU
130, 230, 330, 430, 530	メモリ
140, 240, 340, 440, 540	光通信装置
150, 250, 350, 450, 550	キャッシュ管理装置
160, 260, 360, 460, 560	アービタレション回線
170, 270, 370, 470, 570	データ回線
600	コンセントレータ

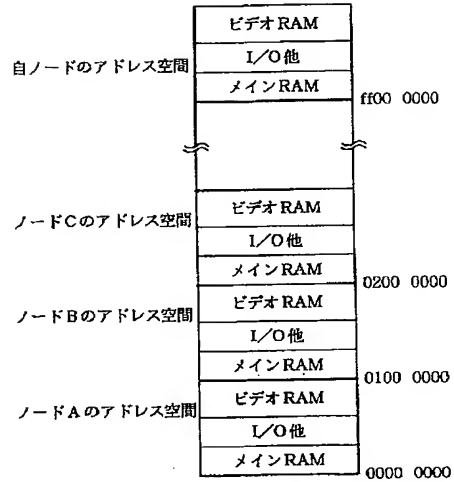
610 光バスアービタ

* * 620 スターカブラ

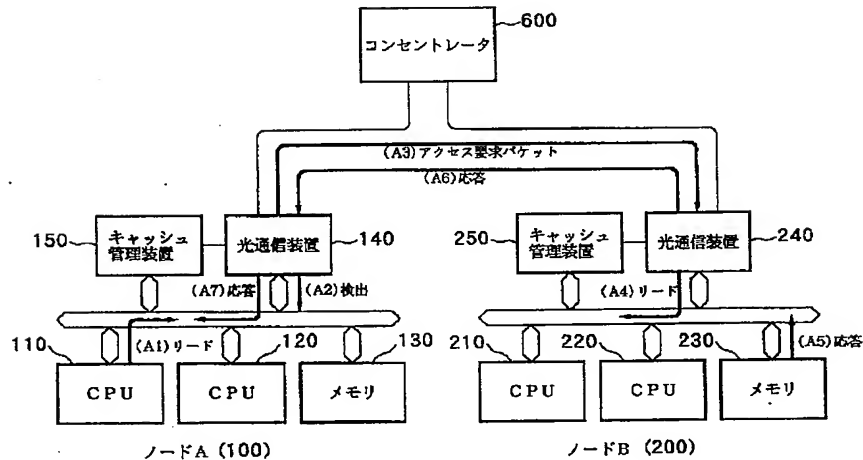
【図1】



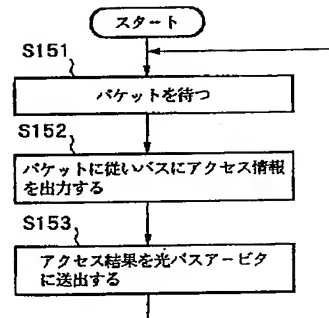
【図2】



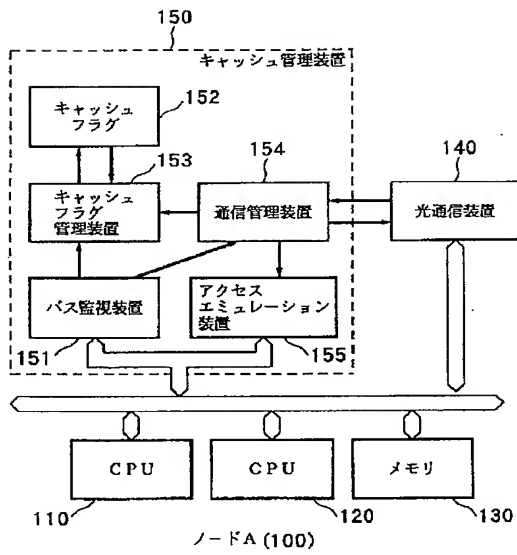
【図3】



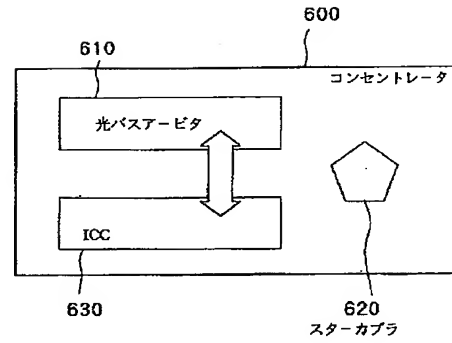
【図11】



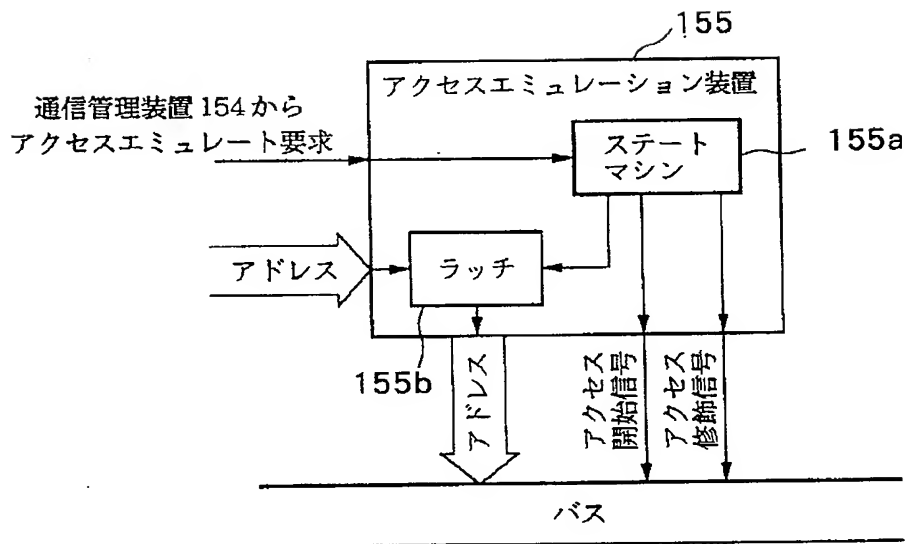
【図4】



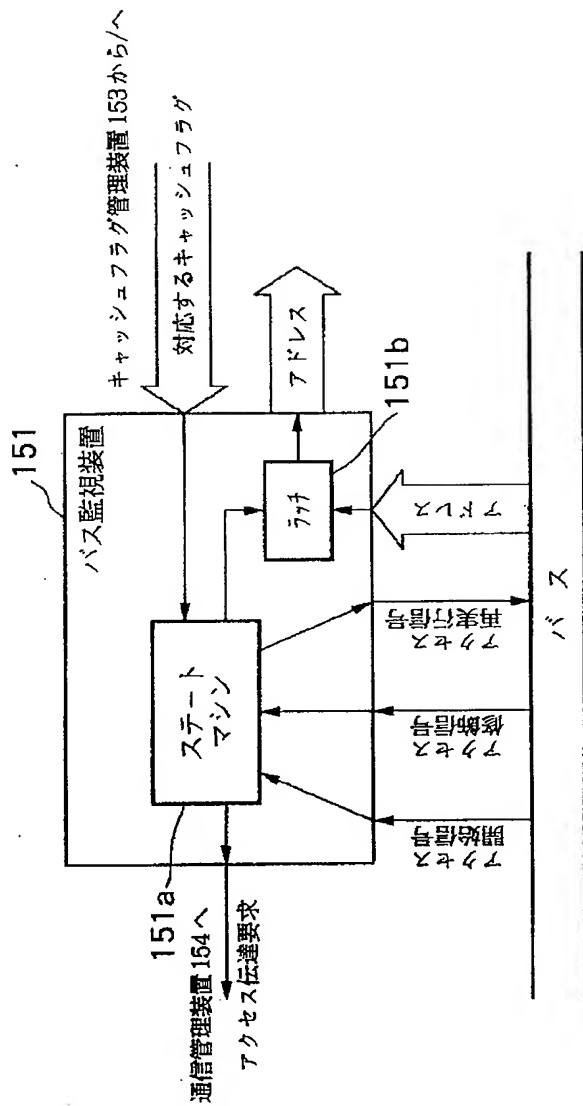
【図8】



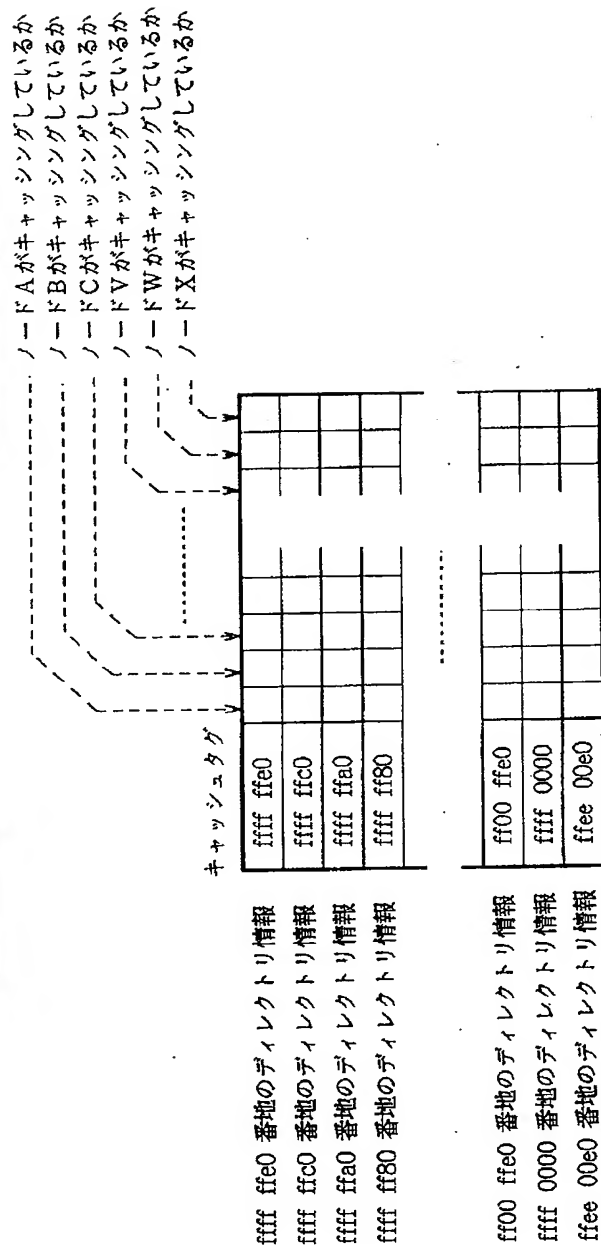
【図7】



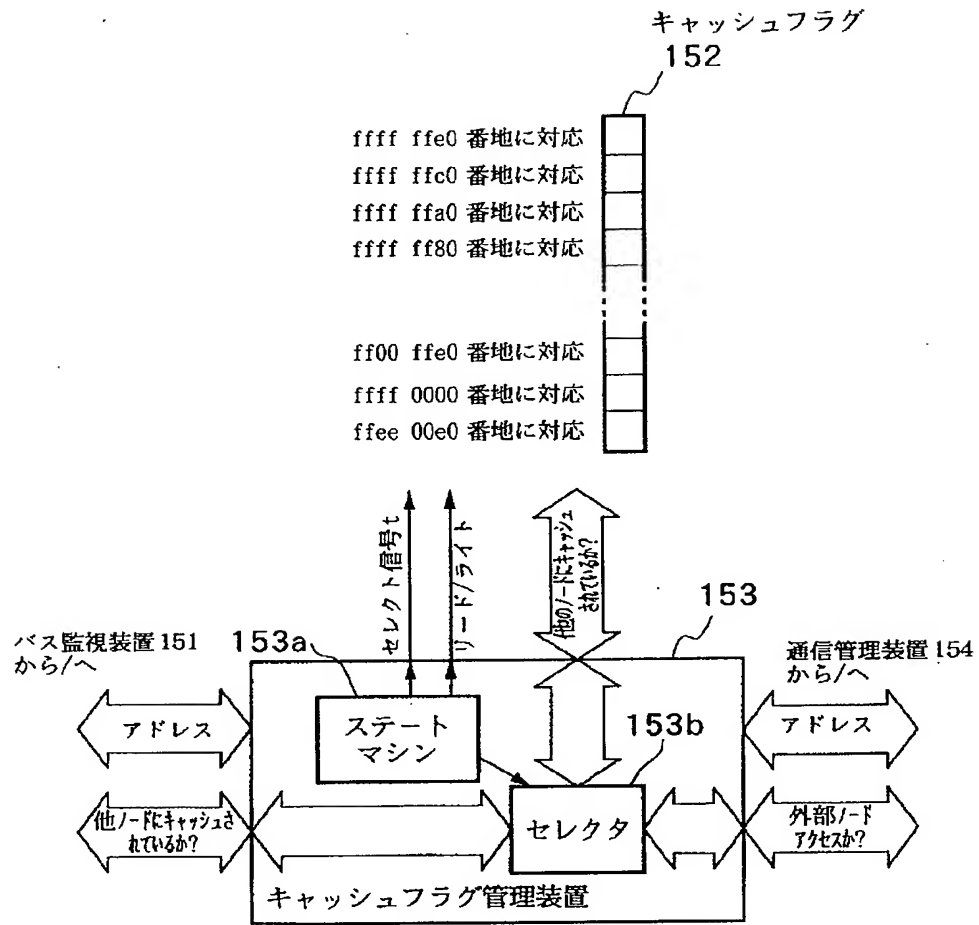
【図5】



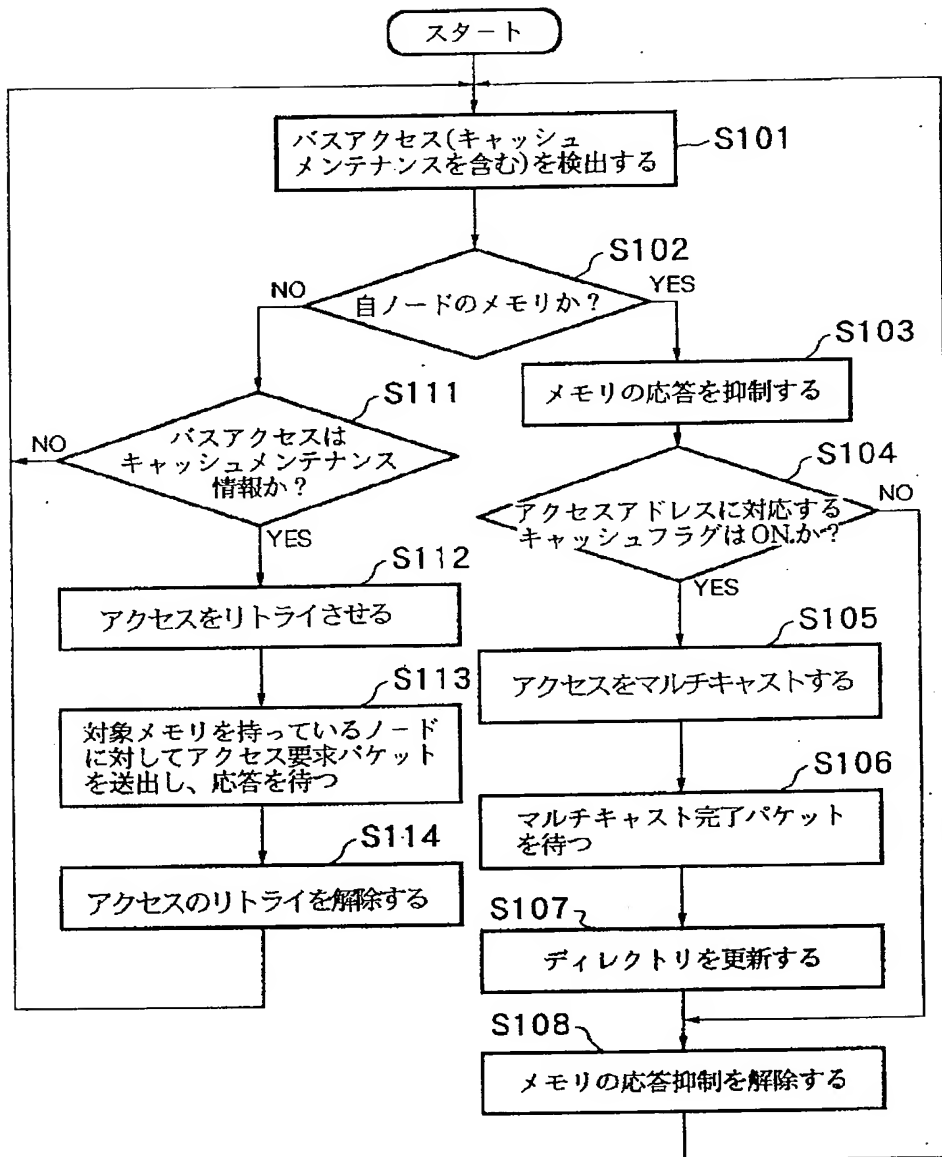
【図9】



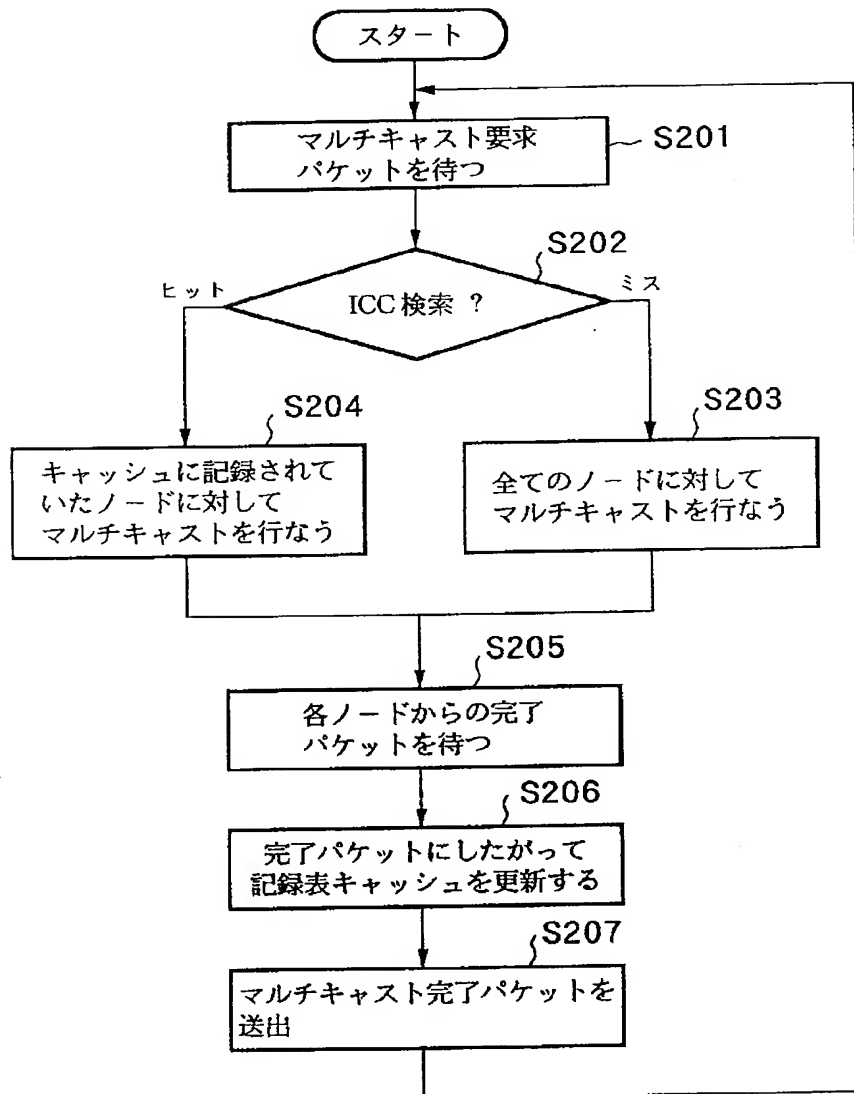
【図6】



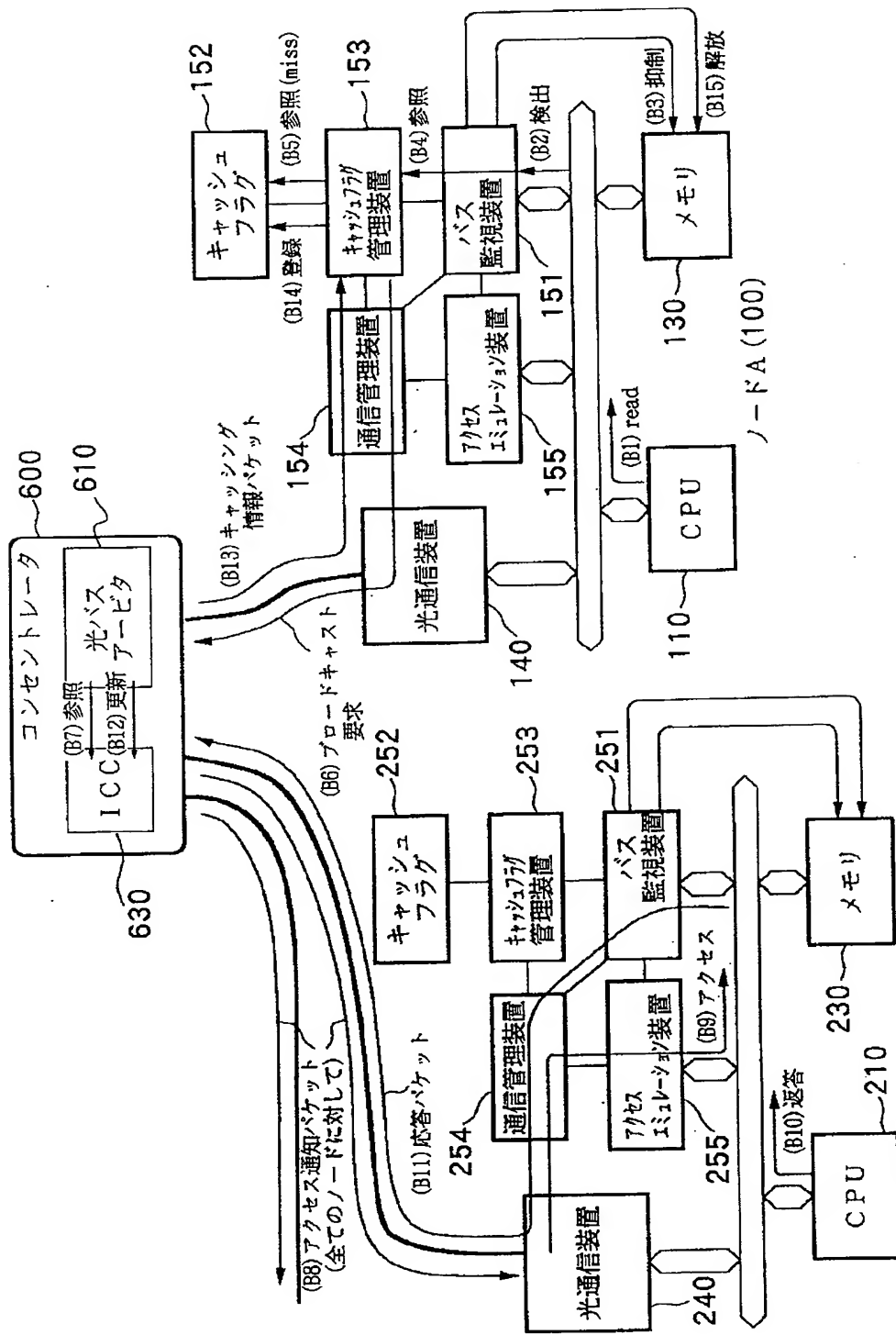
【図10】



【図12】



【図13】



ノードA～Fについて記録していることを示す

ノードID	ノード名	ノード種別	ノード内容	ノードアドレス
f23f ffe0	ノードA	ノードA	f23f ffe0	f23f ffe0
f343 ffc0	ノードB	ノードB	f343 ffc0	f343 ffc0
889f ffa0	ノードC	ノードC	889f ffa0	889f ffa0
f12f ff80	ノードD	ノードD	f12f ff80	f12f ff80

ノードS～Xについて記録していることを示す

ノードID	ノード名	ノードアドレス
0ef0 0040	ノードS	0ef0 0040
0aa5 0020	全ノード	0aa5 0020
003f 6770	ノードH	003f 6770